Modeling The Interplay of People's Location, Interactions, and Social Ties[*]

Adam Sadilek **Henry Kautz** **Jeffrey P. Bigham**
Department of Computer Science
University of Rochester
Rochester, NY 14627
{sadilek,kautz,jbigham}@cs.rochester.edu

Abstract

Location plays an essential role in our lives, bridging our online and offline worlds. This paper explores the interplay of people's location, interactions, and social ties within a large real-world dataset. We present and evaluate Flap, a system that solves two intimately related tasks: link and location prediction in online social networks. For link prediction, Flap infers social ties by considering patterns in friendship formation, the content of people's messages, and user location. We show that while each component is a weak predictor of friendship alone, combining them results in a strong model—accurately identifying the majority of friendships. For location prediction, Flap implements a scalable probabilistic model of human mobility, where we treat users with known GPS positions as noisy sensors of the location of their friends. We explore supervised and unsupervised learning scenarios, and focus on the efficiency of both learning and inference. We evaluate Flap on a large sample of highly active users from two distinct geographical areas and show that it (1) reconstructs the entire friendship graph with high accuracy even when no edges are given; and (2) infers people's fine-grained location, even when they keep their data private and we can only access the location of their friends. Our models significantly outperform current approaches to either task.

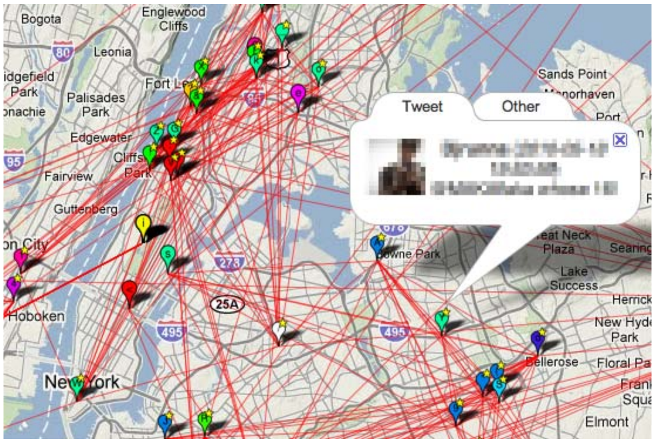

Figure 1: Visualization of a sample of friends in New York City. The red links between users represent friendships, and the colored pins show their current location. This work investigates to what extent can we predict fine-grained location and social ties of individuals on the basis of their online communication.

1 Introduction

Our society is founded on the interplay of human relationships and interactions. Since every person is tightly embedded in our social structure, the vast majority of human behavior can be fully understood only in the context of the actions of others. Thus, not surprisingly, more and more evidence shows that when we want to model the behavior of a person, the best predictors are often not based on the person herself, but rather on her friends, relatives, and other *connected* people. For instance, behavioral patterns of people taking taxis, rating movies, choosing cell phone providers, or sharing music are often best predicted by the habits of related people, rather than by the attributes of the individual such as age, ethnicity, or education [Bell *et al.*, 2007; Pentland, 2008].

Until recently, it was nearly impossible to gather large amounts of data about the connections that play such important roles in our lives. However, this is changing with the explosive increase in the use, popularity, and significance of *online* social media and *mobile* devices.[1] The online aspect makes it practical to collect vast amounts of data, and the mobile element bridges the gap between our online and offline activities. Unlike other computers, phones are aware of the location of their users, and this information is often included in users' posts. In fact, major online social networks are fostering location sharing. Twitter added an explicit GPS tag that can be specified for each tweet (AKA Twitter message update) in early 2010 and is continually im-

[*] This extended abstract is based on the recipient of the best paper award at the Fifth ACM International Conference on Web Search and Data Mining in 2012 [Sadilek *et al.*, 2012a].

[1] http://www.comscore.com/

proving the location-awareness of its service. Google+, Facebook, FourSquare, and Gowalla allow people to share their location, and to "check-in" at venues. With Google Latitude and Bliin, users can *continually* broadcast their location.

As a result, we now have access to colossal amounts of real-world data containing not just the text and images people post, but also their location. Of course, these three data modalities are not necessarily mutually independent. For instance, photos are often GPS-tagged and locations can also be mentioned, or alluded to, in text.

While the information about users' *location* and *relationships* is important to accurately model their behavior and improve their experience, it is not always available. This paper explores novel techniques of inferring this latent information from a stream of message updates (Fig. 1). We present a unified view on the interplay between people's location, message updates, and their social ties on a large real-world dataset. Our approaches are robust and achieve significantly higher accuracy than the best currently known methods, even in difficult experimental settings spanning diverse geographical areas.

2 Significance of Results

Consider the task of determining the exact geographic location of an arbitrary user of an online social network. If she routinely geo-tags her posts and makes them public, the problem is relatively easy. However, suppose the location information is hidden, and you only have access to public posts by her friends. By leveraging social ties, our probabilistic location model—the first component of this work—infers where any given user is with high accuracy and fine granularity in both space and time *even when the user keeps his or her posts private*. Since this work shows that once we have location information for some proportion of people, we can infer the location of their friends, one can imagine doing this recursively until the entire target population is covered. To our knowledge, no other work attempts to predict locations in a comparably difficult setting.

The main power of our link prediction approach—the second major component of this work—is that it accurately reconstructs the entire friendship graph even when no "seed" ties are provided. Previous work either obtained very good predictions at the expense of large computational costs (*e.g.*, [Taskar *et al.*, 2003]), thereby limiting those approaches to very small domains, or sacrificed orders of magnitude in accuracy for tractability (*e.g.*, [Liben-Nowell and Kleinberg, 2007; Crandall *et al.*, 2010]). By contrast, we show that our model's performance is comparable to the most powerful relational methods applied in previous work [Taskar *et al.*, 2003], while at the same time being applicable to large real-world domains with tens of millions (as opposed to hundreds) of possible friendships. Since our model leverages users' locations, it not only encompasses virtual friendships, but also begins to tie them together with their real-life groundings.

Applications of Flap range from improved local content with better social context, through increased security (both personal and electronic) via detection of abnormal behavior tied with one's location, to better organization of one's relationships and connecting virtual friendships with the real-world. Flap can also help contain disease outbreaks [Sadilek *et al.*, 2012b]. Our model allows identification of highly mobile individuals as well as their most likely meeting points, both in the past and in the future. These people can be subsequently selected for targeted treatment or preemptive vaccination. Given people's inferred locations, and limited resource budget, a decision-theoretic approach can be used to select optimal emergency policy. Clearly, strong privacy concerns are tied to such applications, as we discuss in the conclusions.

3 Related Work

Recent research in location-based reasoning has shown that surprisingly rich models of human behavior can be learned from GPS data alone [Liao *et al.*, 2005; Horvitz *et al.*, 2005; Sadilek and Kautz, 2010]. However, previous work has focused solely on relatively small, custom-collected datasets. The content streaming from social networks that we consider here is much richer, more extensive, timely, and interesting—but also more challenging—than the sets of logged location data previously considered.

In the context of online social networks, Cho *et al.* (2011) focus on modeling user location in social networks as a dynamic Gaussian mixture, a generative approach postulating that each check-in is induced from the vicinity of either a person's home, work, or is a result of social influence of one's friends. By contrast, our location model is inherently discrete, which allows us to predict the exact location rather than a sample from a generally high-variance continuous distribution; operates at a finer time granularity; and learns the candidate locations from noisy data. Furthermore, our approach leverages the complex temporal and social dependencies between people's locations in a more general fashion. We show that our model outperforms that of Cho *et al.* in the experiments presented below.

The relationship between social ties and distance has recently received considerable attention [Liben-Nowell *et al.*, 2005; Backstrom *et al.*, 2010; Scellato *et al.*, 2011]. Crandall *et al.* (2010) explore the relationship between co-location of Flickr users and their social ties. They show that the number of distinct places where two users are co-located within various periods of time has the potential to predict a small fraction of the ties quite well. However, the recall is dramatically low. By contrast, with our approach we can predict over 90% of the friendships with confidence beyond 80% (Fig. 2). This is consistent with our experiments, where we show that location alone is generally a poor predictor of friendship. Consider two strangers that happen to take the same train to work, and tweet every time it goes through a station. Our dataset contains a number of instances of this sort.

4 The Data

Our experiments are based on data obtained from Twitter, a popular micro-blogging service where people post at most 140 characters long message updates. The forced brevity encourages frequent mobile updates, as we show below. Relationships between users on Twitter are not necessarily symmetric. One can follow (subscribe to receive messages from) a user without being followed back. When users do reciprocate following, we say they are *friends* on Twitter. There is anecdotal evidence that Twitter friendships have a substantial overlap with offline friendships [Gruzd *et al.*, 2011]. Twitter launched in 2006 and has been experiencing an explosive growth since then. In 2013, over 500 million accounts are registered on Twitter.

Using the Twitter Search API[2], we collected a sample of public tweets that originated from two distinct geographic areas: Los Angeles (LA) and New York City (NYC). The collection period was one month long and started on May 19 2010. We periodically queried Twitter with requests of all recent tweets within 150 kilometers of LA's city center, and 100 kilometers within the NYC city center. Altogether, we have logged over 26 million tweets authored by more than 1.2 million unique users. To put these statistics in context, the entire NYC and LA metropolitan areas have an estimated population of 19 and 13 million people, respectively.[3] In this work, we concentrate on accounts that posted more than 100 GPS-tagged tweets during the one-month data collection period. We refer to them as *geo-active users* and our dataset contains 11,380 of them.

5 The System: Flap

Flap (Friendship + Location Analysis and Prediction), has three main components responsible for downloading Twitter data, visualization (Fig. 1), and learning and inference. We now turn to the third—machine learning—module of Flap that has two main tasks: friendship and location prediction.

5.1 Friendship Prediction

The goal of friendship prediction is to reconstruct the entire social graph, where vertices represent users and edges model friendships. We achieve this via an iterative method that operates over the current graph structure and features of pairs of vertices. We first describe the features used by our model of social ties, and then focus on its structure, learning, and inference. In agreement with prior work, we found that no single property of a pair of individuals is a good indicator of the existence or absence of friendship [Liben-Nowell *et al.*, 2005; Cho *et al.*, 2011]. Therefore, we combine multiple disparate features—based on text (the amount of overlap in the vocabularies of pairs of users), location (the amount of time users spend in close physical proximity), and the topology of the underlying friendship graph (meet/min coefficient of users u and v $\mathcal{M}_{\mathbb{E}}(u,v)$). Since the text and location features scores are always observed, we use a regression decision tree [Breiman and others, 1984] to unify them into one feature $\mathcal{DT}(u,v)$. Thus, we end up with one feature function for the observed variables ($\mathcal{DT}$) and one for the hidden variables ($\mathcal{M}_{\mathbb{E}}$).

Our probabilistic model of the friendship network is a Markov random field that has a hidden node for each possible friendship. Since the friendship relationship is symmetric and irreflexive, our model contains $n(n-1)/2$ hidden nodes, where n is the number of users. Each hidden node is connected to an observed node ($\mathcal{DT}$) and to all other hidden nodes. Ultimately, we are interested in the probability of existence of an edge (friendship) given the current graph structure and the pairwise features of the vertices (users) the edge is incident on. Applying Bayes' theorem while assuming mutual independence of features $\mathcal{DT}$ and $\mathcal{M}_{\mathbb{E}}$, we can write

$$\begin{aligned} &P(E=1|DT=d, M_{\mathbb{E}}=m) \propto \\ &= P(DT=d|E=1)P(E=1|M_{\mathbb{E}}=m)/Z \end{aligned} \tag{1}$$

We evaluate Flap on friendship prediction using two-fold cross-validation in which we train on LA and test on NY data, and vice versa. We average the results over the two runs. We varied the amount of randomly selected edges provided to the model at testing time from 0 to 50%.

Flap reconstructs the friendship graph well over a wide range of conditions—even when given no edges (Fig. 2). It far outperforms the baseline model (decision tree) and the precision/recall breakeven points are comparable to those of [Taskar *et al.*, 2003], even though our domain is orders of magnitude larger and our model is more tractable.

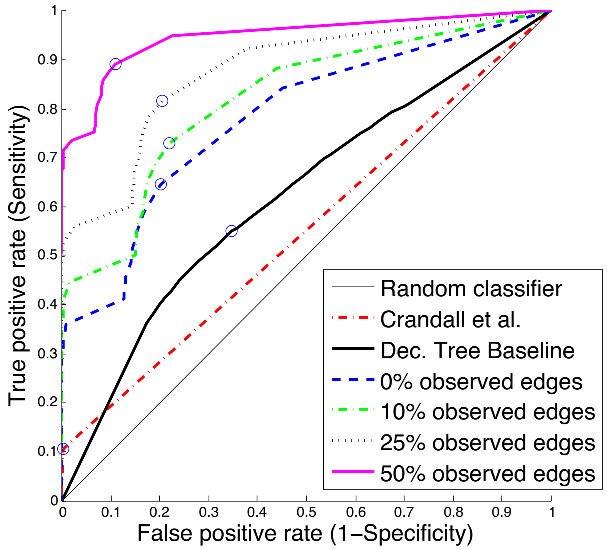

Figure 2: Averaged ROC curves for decision tree baseline, Crandall *et al.*'s model with the most favorable setting of parameters ($s = 0.001$ and $t = 4$ hours), and Flap.

At learning time, we first train a regression decision tree $\mathcal{DT}$ and prune it using ten-fold cross-validation to prevent overfitting. We also perform maximum likelihood learning of the parameters $P(DT|E)$ and $P(E|M_{\mathbb{E}})$.

At inference time, we use $\mathcal{DT}$ to make preliminary predictions on the test data. Next, we execute a customized loopy belief propagation algorithm that is initialized with the probabilities estimated by $\mathcal{DT}$. Even though the graphical model is dense, our algorithm converges within several hundred iterations, due in part to the sufficiently accurate initialization and regularization provided by the decision tree.

5.2 Location Prediction

The goal of Flap's location prediction component is to infer the most likely location of person u at any time. The input

[2] http://search.twitter.com/api/

[3] http://www.census.gov/popest/metro/

consists of a sequence of locations visited by u's friends (and for supervised learning, locations of u himself over the training period), along with corresponding time information. The model outputs the most likely sequence of locations u visited over a given time period.

We model user location in a dynamic Bayesian network [Murphy, 2002]. In each time slice, we have one hidden node and a number of observed nodes, all of which are discrete. The hidden node represents the location of the target user (u). The node td represents the time of day and w determines if a given day is a work day or a free day (weekend or a national holiday). Each of the remaining observed nodes ($f1$ through fn) represents the location of one of the target user's friends. Since the average node degree of geo-active users is 9.2, we concentrate on $n \in \{0, 1, 2, \ldots, 9\}$, although our approach works for arbitrary nonnegative values of n. Each node is indexed by time slice. We model each person's discrete location in 20 minute increments, since more than 90% of the users tweet with lower frequency [Sadilek *et al.*, 2012a].

We explore both supervised and unsupervised learning of user mobility. In the earlier case, for each user, we train a DBN on the first three weeks of data with known hidden location values. In the latter case, the hidden labels are unknown to the system.

At inference time, we are interested in the most likely explanation of the observed data. That is, given a sequence of locations visited by one's friends, along with the corresponding time and day type, our model outputs the most likely sequence of locations one visited over the given time period.

Our evaluation is done in terms of *accuracy*—the percentage of timeslices for which the model infers the correct user location. We have a separate dynamic Bayesian network model for each user. We evaluate the overall performance via cross-validation. In each fold of cross-validation, we designate a target user and run learning and inference for him. This process is repeated for all users, and we report the average results over all runs for a given value of n.

Our supervised DBN model predicts people's exact location with up to 84% accuracy, whereas the unsupervised version yields up to 57% accuracy. The accuracy of most-frequent baseline is 28%. As n increases, the accuracy improves. However, information about the top-2 most active friends provides the largest boost in accuracy (for $n > 2$ the accuracy curves plateau).

6 Conclusions and Future Work

Location information linked with the content of users' messages in online social networks is a rich information source that is now accessible to machines in massive volumes and at ever-increasing real-time streaming rates. In this work, we show that there are significant patterns that characterize locations of individuals and their friends. These patterns can be leveraged in probabilistic models that infer people's locations as well as social ties with high accuracy. Moreover, the prediction accuracy degrades gracefully as we limit the amount of observed data available to the models, suggesting successful future deployment of Flap at a scale of an entire social network.

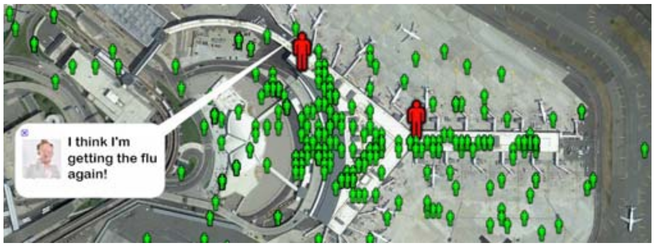

Figure 3: Visualization of a sample of Twitter users at an airport. Individuals who indicate sickness in their Twitter feed are highlighted in red. From the GPS-tagged data, we see who likely came into contact with the infected people. Additional candidates can be inferred using methods presented in this paper. Emerging research is already beginning to show that putting all this information together yields stronger *predictions* about the spread of a contagious disease which could lead to better disease prevention and containment.

By training the model on one geographical area and testing on the other using cross-validation (total of 4 million geotagged public tweets we collected from Los Angeles and New York City metropolitan areas), we show that Flap discovers robust patterns in the formation of friendships that transcend diverse and distant areas of the USA. We conclude that no single property of a pair of individuals is a good indicator of the existence or absence of friendship. And no single friend is a good predictor of one's location. Rather, we need to combine multiple disparate features—based on text, location, and the topology of the underlying friendship graph—in order to achieve good performance.

We recognize that there are substantial ethical questions ahead, specifically concerning tradeoffs between the values our automated systems create versus user privacy. For example, our unsupervised experiments show that location can be inferred even for people who keep their tweets and location private, and thus may believe that they are "untrackable." These issues will need to be addressed in parallel with the development of AI models.

However, we believe that the benefits of Flap—in helping to connect and localize users, and in building smarter systems—outweigh the possible dangers. There are many exciting practical applications that have the potential to change people's lives that rely on location and link prediction. For example, we have recently shown that that geo-tagged social media posts can be used to to predict the onset of flu-like disease in specific individuals and measure behavioral influence on illness [Sadilek *et al.*, 2012c]. This and similar work on low-cost, real-time, population-scale health monitoring using social media could revolutionize public health (Fig. 3).

7 Acknowledgments

This research was partly funded by ARO grant W911NF-08-1-0242, ONR grant N00014-11-10417, OSD grant W81XWH-08-C0740, and a gift from the Kodak Company.

References

[Backstrom *et al.*, 2010] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

[Bell *et al.*, 2007] Robert Bell, Yehuda Koren, and Chris Volinsky. Modeling relationships at multiple scales to improve accuracy of large recommender systems. In *KDD*, pages 95–104, New York, NY, USA, 2007. ACM.

[Breiman and others, 1984] Leo Breiman et al. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

[Cho *et al.*, 2011] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2011.

[Crandall *et al.*, 2010] D.J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436, 2010.

[Gruzd *et al.*, 2011] Anatoliy Gruzd, Barry Wellman, and Yuri Takhteyev. Imagining Twitter as an imagined community. In *American Behavioral Scientist, Special issue on Imagined Communities*, 2011.

[Horvitz *et al.*, 2005] E. Horvitz, J. Apacible, R. Sarin, and L. Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *Twenty-First Conference on Uncertainty in Artificial Intelligence*, 2005.

[Liao *et al.*, 2005] Lin Liao, Dieter Fox, and Henry Kautz. Location-based activity recognition using relational Markov networks. In *IJCAI*, 2005.

[Liben-Nowell and Kleinberg, 2007] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58:1019–1031, May 2007.

[Liben-Nowell *et al.*, 2005] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623, 2005.

[Murphy, 2002] Kevin P. Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

[Pentland, 2008] Alex (Sandy) Pentland. *Honest Signals: How They Shape Our World*. The MIT Press, 2008.

[Sadilek and Kautz, 2010] Adam Sadilek and Henry Kautz. Recognizing multi-agent activities from GPS data. In *Twent-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[Sadilek *et al.*, 2012a] Adam Sadilek, Henry Kautz, and Jeffrey P. Bigham. Finding your friends and following them to where you are. In *Fifth ACM International Conference on Web Search and Data Mining*, 2012. (Best Paper Award).

[Sadilek *et al.*, 2012b] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Modeling spread of disease from social interactions. In *Sixth AAAI International Conference on Weblogs and Social Media (ICWSM)*, 2012.

[Sadilek *et al.*, 2012c] Adam Sadilek, Henry Kautz, and Vincent Silenzio. Predicting disease transmission from geo-tagged micro-blog data. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[Scellato *et al.*, 2011] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of ICWSM*, 11, 2011.

[Taskar *et al.*, 2003] Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller. Link prediction in relational data. In *in Neural Information Processing Systems*, 2003.